D207 Exploratory Data Analysis

Nicole Reiswig

College of Information Technology, Western Governors University

David Gagner

August 31, 2023

D207 Exploratory Data Analysis

A.

1. The question that will be researched and answered in the exploratory data analysis is, “What effect does Timely_Response have on Tenure?”

2. Stakeholders will benefit from the answer to this question. The stakeholders that will benefit from this are the customers, investors, employees, and the company as a whole. The customer will benefit from this question because we can identify the areas where the business excels, the weak points, and what effects it has on customers' length of tenure. Once we are aware of the relevant factors, the business can continue performing the positives well and improve the areas that are not performing well to increase customer tenure and satisfaction. This will result in service satisfaction from the customers. Once the business acts on the data acquired during this analysis employees will benefit as they will be given the resources to do their jobs well. Once the employees have the proper resources to satisfy the customers, they will have higher satisfaction with their jobs and retain employment. The business will benefit as they will know what their customers want and need to continue doing business long-term with the company, which results in steady cash flow. The investors will benefit from this analysis because as the company makes these improvements, the cash flow becomes more stable, and turnover reduces, which means positive cash flow for the investors.

3. All the variables in the data set have the potential to provide relevant data to analyze. We will utilize the variables Tenure and Timely_Response to answer the research question. We will also review Timely_Fixes and Income. Tenure is a quantitative continuous variable and is described as the number of months the customer has stayed with the provider. Timely_Response is categorical ordinal variable which is one of eight questions sent out to customers on a survey in

which customers rank the importance of these factors on a scale of 1 to 8 with 1 being most important and 8 being least important.

B.

1. The technique that is utilized to answer the research question will be the ANOVA technique. We will be analyzing a quantitative continuous variable Tenure and a categorical ordinal variable Timely_Response. ANOVA is a comparison test in which the effect a categorical variable has on the mean of some other characteristic. ANOVA was the test of choice because we are testing a categorical variable and a numerical variable against one another.

```r
#Check working directory
getwd()

#data profiling
str("~/MSDA/churn_clean.csv")
str("C:/Users/ntrei/OneDrive/Documents/MSDA/churn_clean.csv")

#dimension of churn_clean [in-text citation: (R programming 101, n.d.)]
dim(churn_clean)
library(tidyverse)
glimpse(churn_clean)

#rename column names item 1- 8 [in-text citation: (Zach, 2022)]
colnames(churn_clean)[colnames(churn_clean) == 'Item1'] <- 'Timely_Response'
colnames(churn_clean)[colnames(churn_clean) == 'Item2'] <- 'Timely_Fixes'
colnames(churn_clean)[colnames(churn_clean) == 'Item3'] <- 'Timely_Replacements'
colnames(churn_clean)[colnames(churn_clean) == 'Item4'] <- 'Reliability'
colnames(churn_clean)[colnames(churn_clean) == 'Item5'] <- 'Options'
colnames(churn_clean)[colnames(churn_clean) == 'Item6'] <- 'Respectful_Response'
colnames(churn_clean)[colnames(churn_clean) == 'Item7'] <- 'Courteous_Exchange'
colnames(churn_clean)[colnames(churn_clean) == 'Item8'] <- 'Active_Listening'

```

```
> glimpse(churn_clean)
Rows: 10,000
Columns: 50
$ CaseOrder            <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 1…
$ Customer_id          <chr> "K409198", "S120509", "K191035", "D90850", "K662701", "W303516",…
$ Interaction          <chr> "aa90260b-4141-4a24-8e36-b04ce1f4f77b", "fb76459f-c047-4a9d-8af9…
$ UID                  <chr> "e885b299883d4f9fb18e39c75155d990", "f2de8bef964785f41a295982983…
$ City                 <chr> "Point Baker", "West Branch", "Yamhill", "Del Mar", "Needville",…
$ State                <chr> "AK", "MI", "OR", "CA", "TX", "GA", "TN", "OK", "FL", "OH", "PA"…
$ County               <chr> "Prince of Wales-Hyder", "Ogemaw", "Yamhill", "San Diego", "Fort…
$ Zip                  <int> 99927, 48661, 97148, 92014, 77461, 31030, 37847, 73109, 34771, 4…
$ Lat                  <dbl> 56.25100, 44.32893, 45.35589, 32.96687, 29.38012, 32.57032, 36.4…
$ Lng                  <dbl> -133.37571, -84.24080, -123.24657, -117.24798, -95.80673, -83.89…
$ Population           <int> 38, 10446, 3735, 13863, 11352, 17701, 2535, 23144, 17351, 20193,…
$ Area                 <chr> "Urban", "Urban", "Urban", "Suburban", "Suburban", "Urban", "Sub…
$ Timely_Response      <int> 5, 3, 4, 4, 4, 3, 6, 2, 5, 2, 4, 4, 1, 5, 3, 3, 3, 2, 3, 5, 2, 5…
$ Timely_Fixes         <int> 5, 4, 4, 4, 4, 3, 5, 2, 4, 2, 4, 4, 2, 6, 3, 3, 4, 2, 4, 5, 3, 5…
$ Timely_Replacements  <int> 5, 3, 2, 4, 4, 3, 6, 2, 4, 2, 4, 3, 1, 5, 4, 3, 4, 4, 3, 5, 3, 4…
$ Reliability          <int> 3, 3, 4, 2, 3, 2, 4, 5, 3, 2, 7, 4, 4, 2, 2, 2, 3, 3, 4, 4, 2, 4…
$ Options              <int> 4, 4, 4, 5, 4, 4, 1, 2, 4, 5, 3, 4, 3, 4, 3, 4, 5, 5, 3, 3, 4, 4, 3…
$ Respectful_Response  <int> 4, 3, 3, 4, 4, 3, 5, 3, 3, 2, 3, 4, 2, 5, 4, 3, 4, 4, 2, 4, 2, 4…
$ Courteous_Exchange   <int> 3, 4, 3, 3, 4, 3, 5, 4, 4, 3, 3, 3, 3, 4, 4, 5, 4, 3, 3, 3, 3, 3…
$ Active_Listening     <int> 4, 4, 3, 3, 5, 3, 5, 5, 4, 3, 4, 4, 3, 4, 2, 2, 3, 3, 5, 3, 3, 5…
> 8
```

```
22
23  #Verify columns were re-named successfully
24  glimpse(churn_clean)
25
26  #explore data variables [in-text citation: (R programming 101, n.d.)]
27  View(sort(table(churn_clean$Tenure)))
28  View(sort(table(churn_clean$Churn)))
29  View(sort(table(churn_clean$Timely_Response)))
30  View(sort(table(churn_clean$Timely_Fixes)))
31
32  barplot(sort(table(churn_clean$Churn)))
33  barplot(sort(table(churn_clean$Tenure)))
34  barplot(sort(table(churn_clean$Timely_Response)))
35  barplot(sort(table(churn_clean$Timely_Fixes)))
36
37  mean(churn_clean$Tenure)
38  mean(churn_clean$Timely_Response)
39  mean(churn_clean$Timely_Fixes)
40
41  hist(churn_clean$Tenure)
42  hist(churn_clean$Timely_Response)
43  hist(churn_clean$Timely_Fixes)
```

Untitled1* Untitled2* sort(table(churn_clean$Tenure))

Filter

	Var1	Freq
1	1.00025934	1
2	1.005104	1
3	1.018519598	1
4	1.019306038	1
5	1.025579	1
6	1.030630574	1
7	1.033067819	1
8	1.034561	1
9	1.042140757	1
10	1.042724425	1
11	1.047382039	1
12	1.050155	1
13	1.051188042	1
14	1.059865	1

Showing 1 to 15 of 9,996 entries, 2 total columns

Untitled1* Untitled2* sort(table(churn_clean$Timely_Respons...

Filter

	Var1	Freq
1	7	19
2	6	199
3	1	224
4	5	1359
5	2	1393
6	4	3358
7	3	3448

```
> mean(churn_clean$Tenure)
[1] 34.52619
> mean(churn_clean$Timely_Response)
[1] 3.4908
> 8
```

```
44
45 b <- boxplot(churn_clean$Tenure)
46 b <- boxplot(churn_clean$Timely_Response)
47 b <- boxplot(churn_clean$Timely_Fixes)
48
49 churn_clean %>%
50   summarise(mean(Timely_Response), mean(Timely_Fixes), mean(Tenure))
51
52 churn_clean %>%
53   summarise(median(Timely_Response), median(Timely_Fixes), median(Tenure))
54
55 #Create a table with data columns
56 My_data <- churn_clean %>%
57   select(Tenure, Timely_Response, Timely_Fixes) %>%
58   drop_na()
59
```

```
+   summarise(mean(Timely_Response), mean(Timely_Fixes), mean(Tenure))
  mean(Timely_Response) mean(Timely_Fixes) mean(Tenure)
1               3.4908             3.5051     34.52619
> 8
```

```
+   summarise(median(Timely_Response), median(Timely_Fixes), median(Tenure))
  median(Timely_Response) median(Timely_Fixes) median(Tenure)
1                       3                    4       35.43051
> 8
```

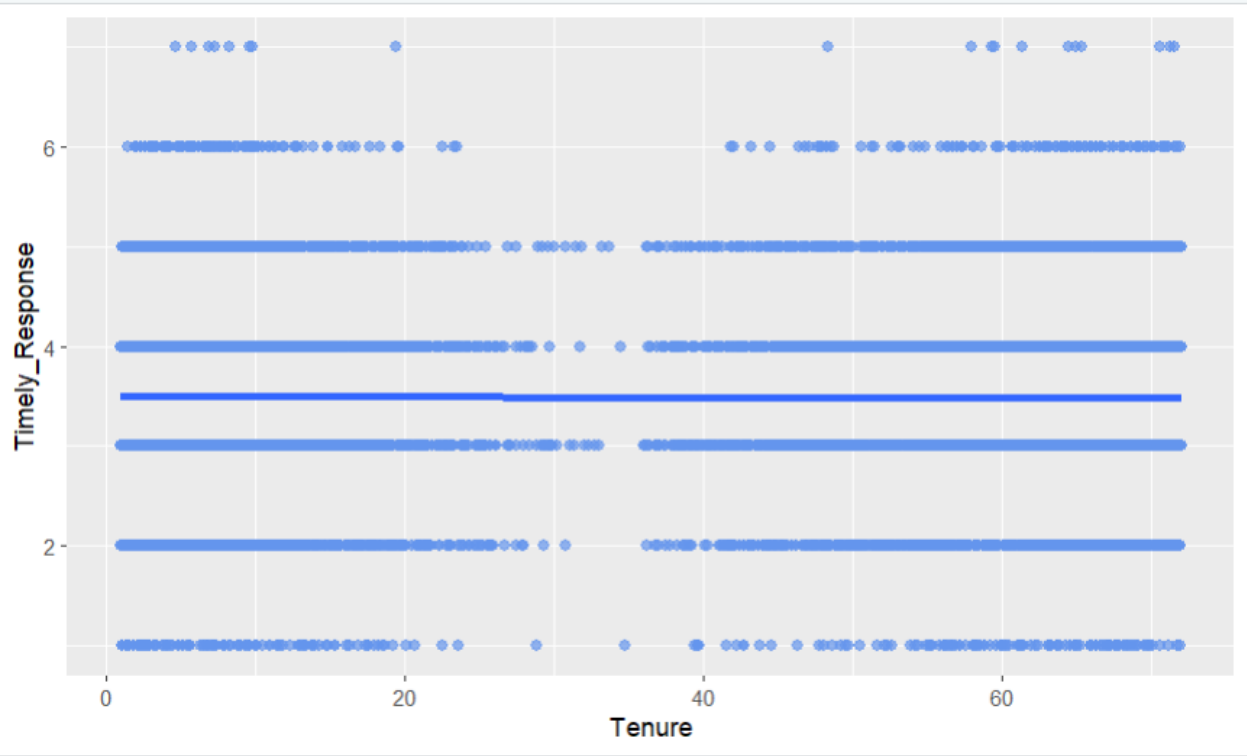
Untitled1* Untitled2* My_data

Filter

	Tenure	Timely_Response	Timely_Fixes
1	6.795513	5	5
2	1.156681	3	4
3	15.754144	4	4
4	17.087227	4	4
5	1.670972	4	4
6	7.000994	3	3
7	13.236774	6	5
8	4.264255	2	2
9	8.220686	5	4
10	3.422086	2	2
11	19.267262	4	4
12	10.521998	4	4
13	13.011492	1	2
14	16.879220	5	6

Showing 1 to 15 of 10,000 entries, 3 total columns

```
59
60 #Graph data
61 library(ggplot2)
62   ggplot(data = My_data,
63   mapping =  aes(x = Tenure, y = Timely_Response)) +
64     geom_point(color = "cornflowerblue",
65                alpha = .7,
66                size = 2) +
67     geom_smooth(method = "lm",
68                 se = FALSE,
69                 linewidth = 1.5)
70
71
```

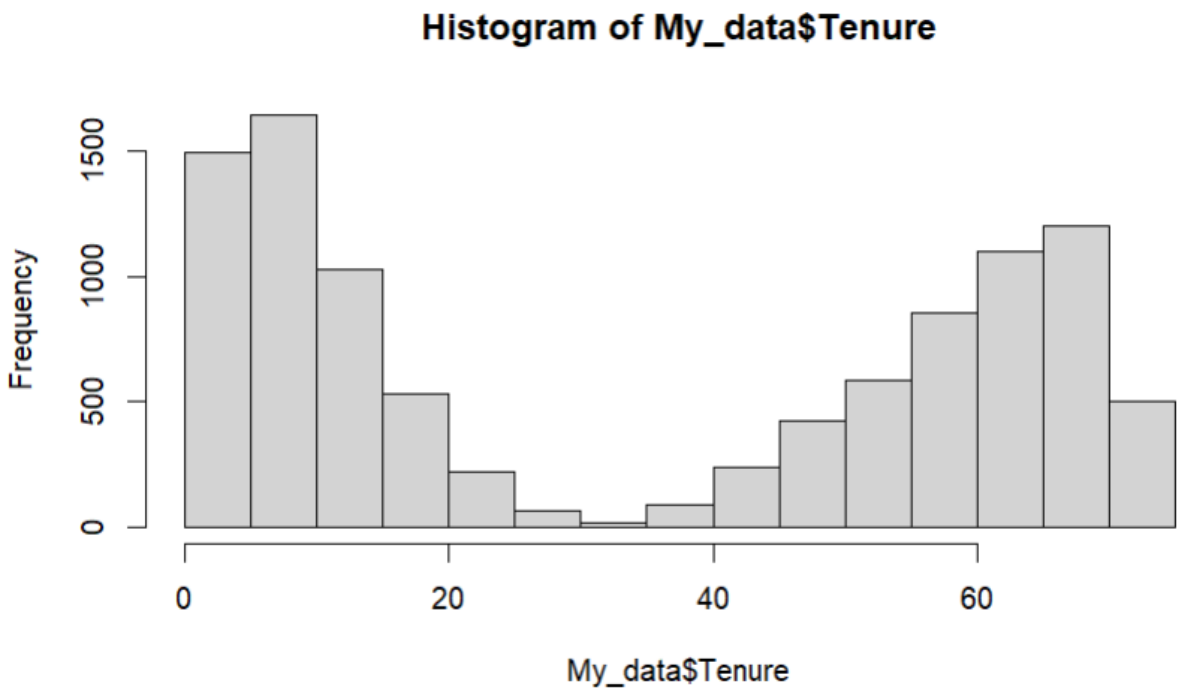


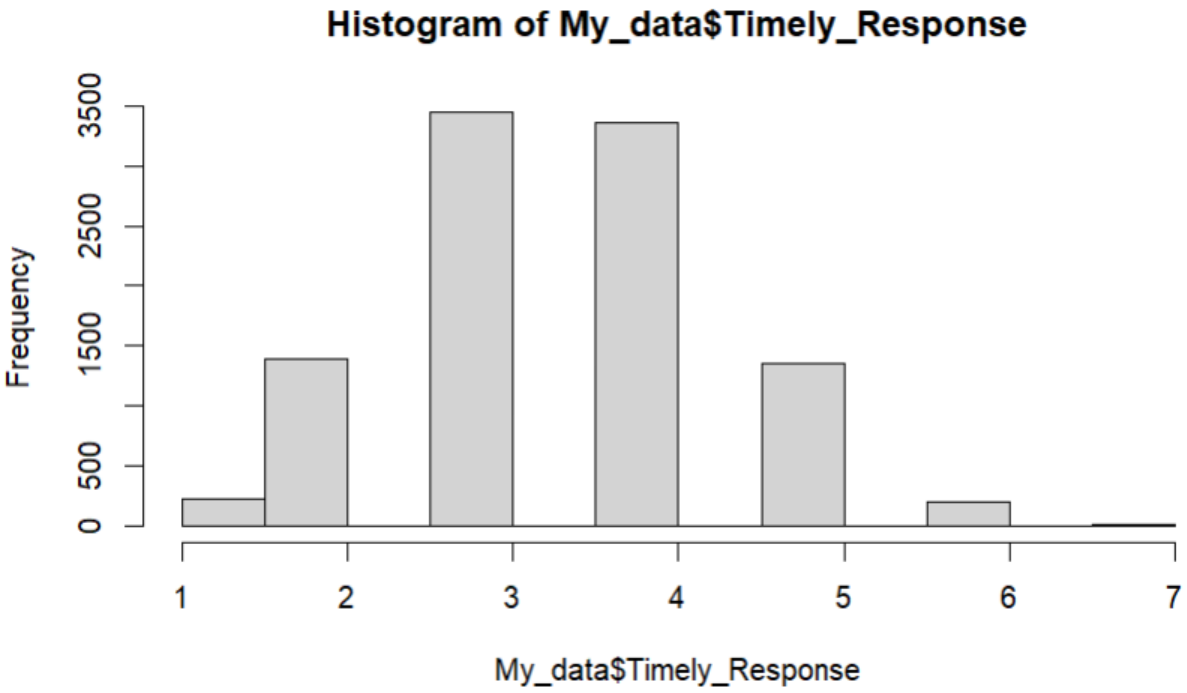

```
71
72  #ANOVA (R programming 101, n.d.)]
73  #Research question: What is the effect of Timely_Response on the length of tenure?
74  #Hypothesis: Ho- there is a correlation between the mean tenure and the survey
75  #question  H1- there is not a correlation between the mean tenure and the survey question
76  #Calculate test statistic
77
78  My_data %>%
79    select(Tenure, Timely_Response, Timely_Fixes) %>%
80    drop_na()
81
82  My_anova <-
83    aov(My_data$Timely_Response ~ My_data$Tenure, data = My_data)
84
85  summary(my_anova)
86
87  My_anova2 <- aov(My_data$Timely_Fixes ~ My_data$Tenure, data = My_data)
88
89  summary(My_anova2)
90
91  #univariate graphs
92  hist(My_data$Tenure)
93  hist(My_data$Timely_Response)
```

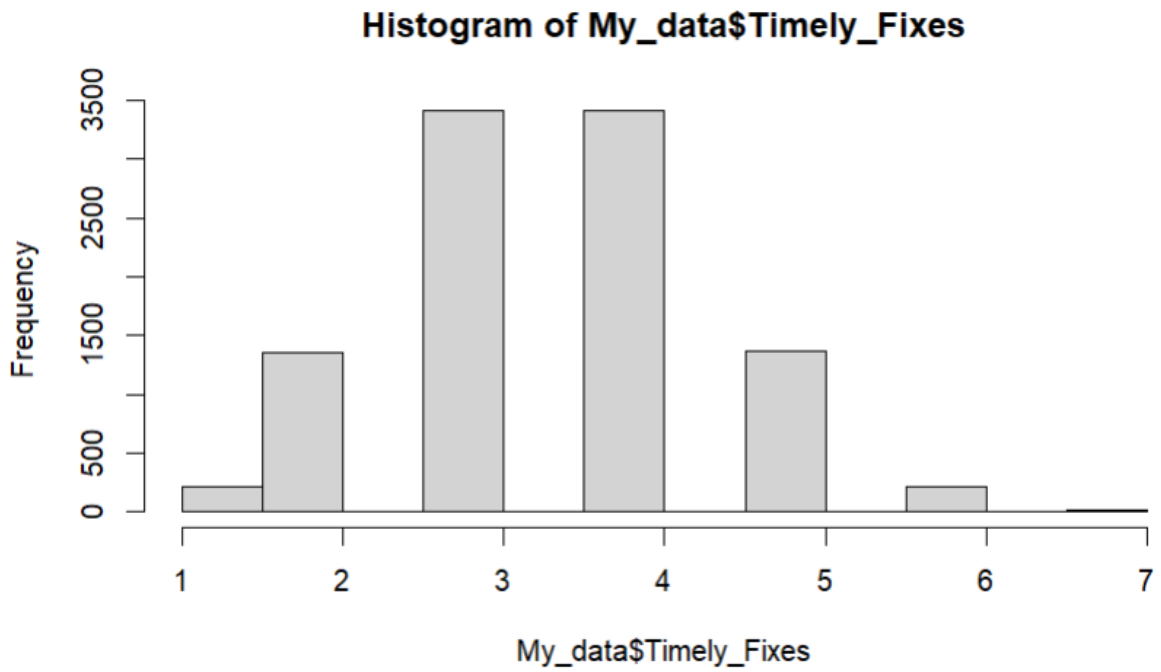
```
> My_data %>%
+   select(Tenure, Timely_Response, Timely_Fixes) %>%
+   drop_na()
       Tenure Timely_Response Timely_Fixes
1    6.795513                5           5
2    1.156681                3           4
3   15.754144                4           4
4   17.087227                4           4
5    1.670972                4           4
6    7.000994                3           3
7   13.236774                6           5
8    4.264255                2           2
9    8.220686                5           4
10   3.422086                2           2
11  19.267262                4           4
12  10.521998                4           4
```

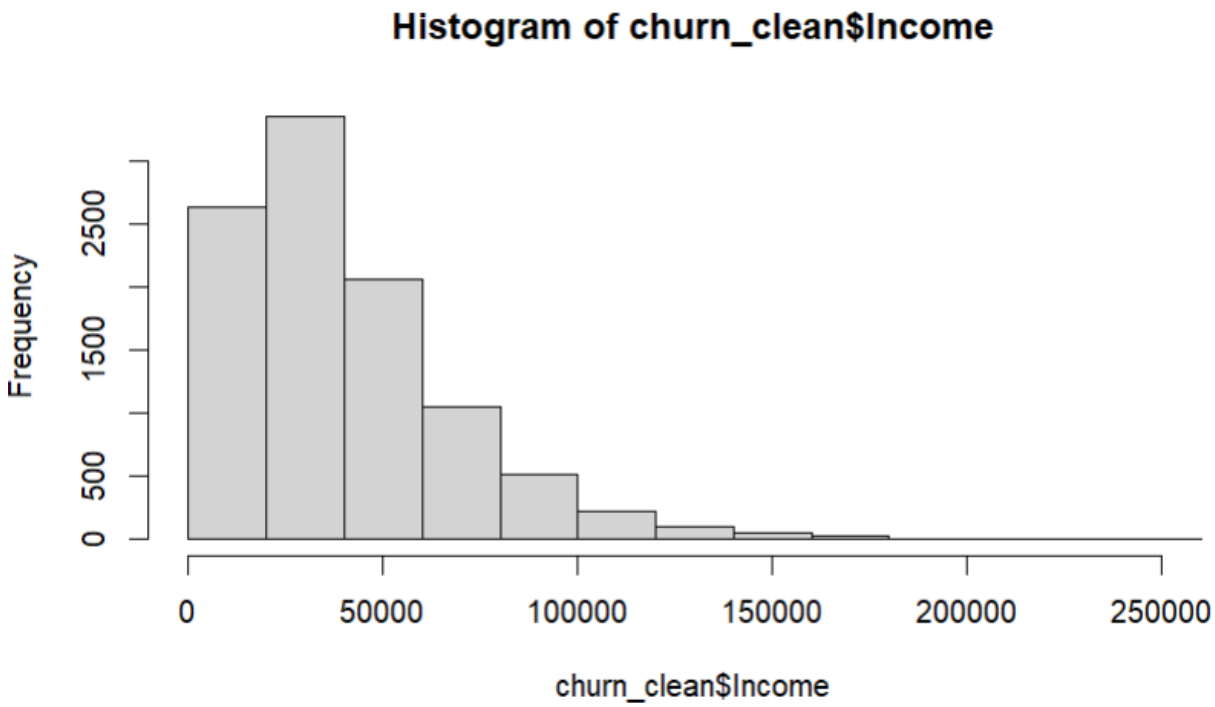
```
> My_anova <-
+   aov(My_data$Timely_Response ~ My_data$Tenure, data = My_data)
> summary(my_anova)
                    Df Sum Sq Mean Sq F value Pr(>F)
churn_clean$Tenure   1      0  0.4201    0.39  0.532
Residuals         9998  10769  1.0771
> 8
```

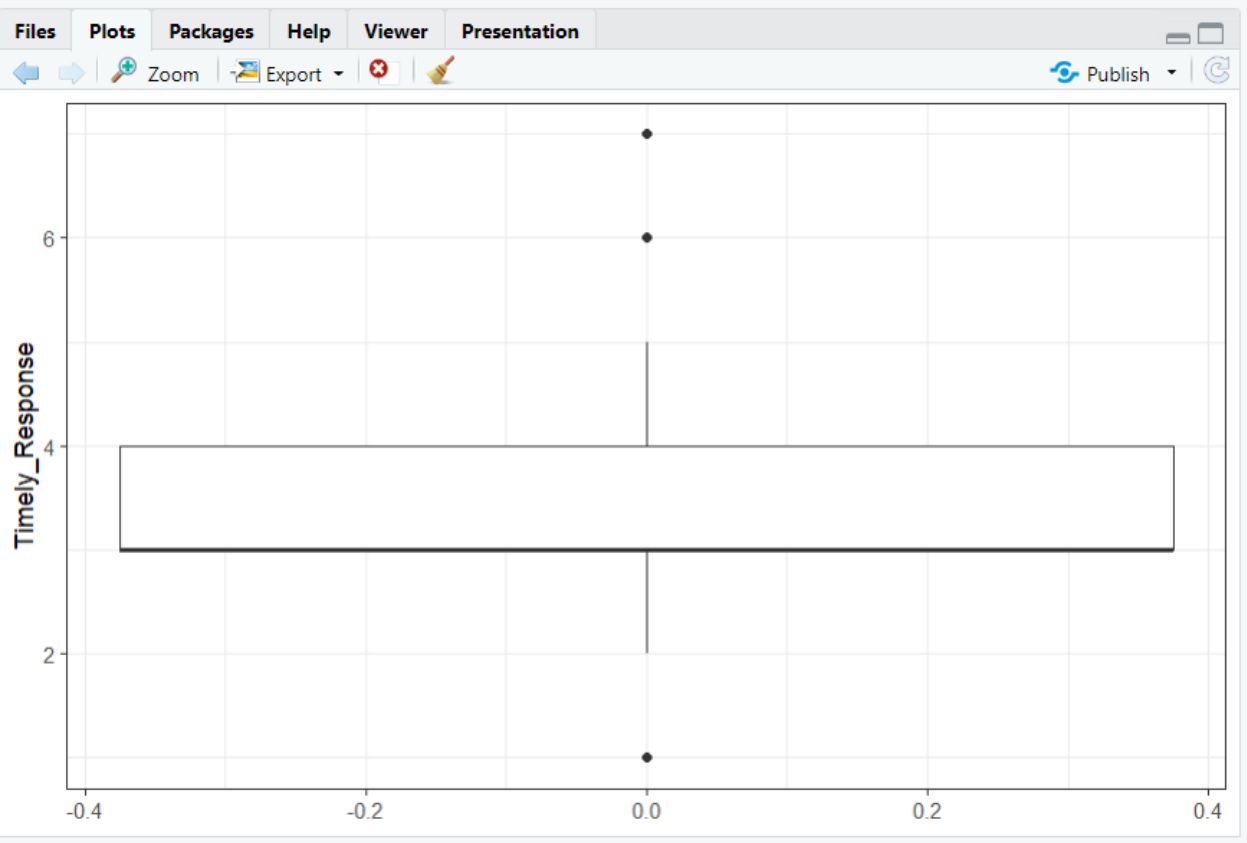
```
91  #univariate graphs
92  hist(My_data$Tenure)
93  hist(My_data$Timely_Response)
94  hist(My_data$Timely_Fixes)
95  hist(churn_clean$Income)
96
97
98  My_data %>%
99    drop_na(Timely_Response) %>%
100   ggplot(aes(Timely_Response)) +
101   geom_boxplot() +
102   coord_flip() +
103   theme_bw()
104
105 My_data %>%
106   drop_na(Timely_Fixes) %>%
107   ggplot(aes(Timely_Fixes)) +
108   geom_boxplot() +
109   coord_flip() +
110   theme_bw()
111
```

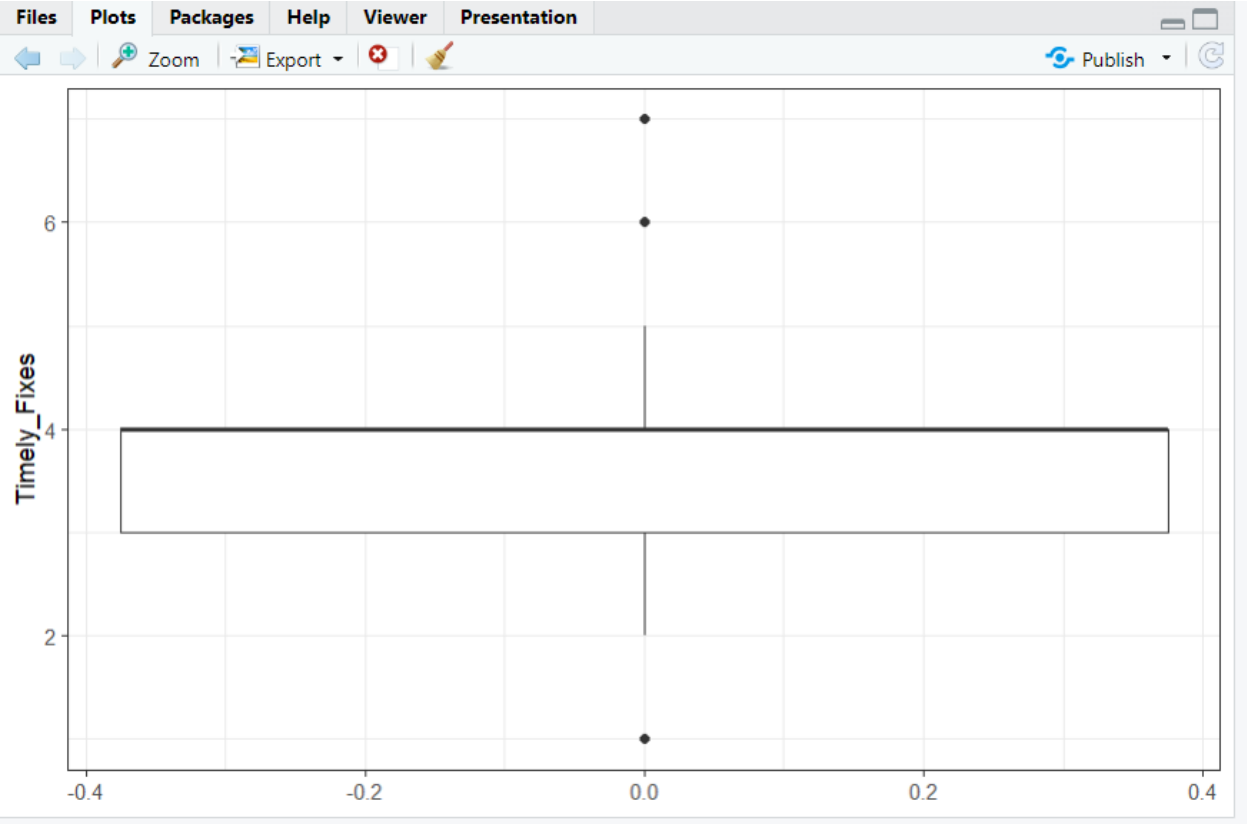

Histogram of My_data$Tenure
Frequency
0
500
1000
1500
0
20
40
60
My_data$Tenure

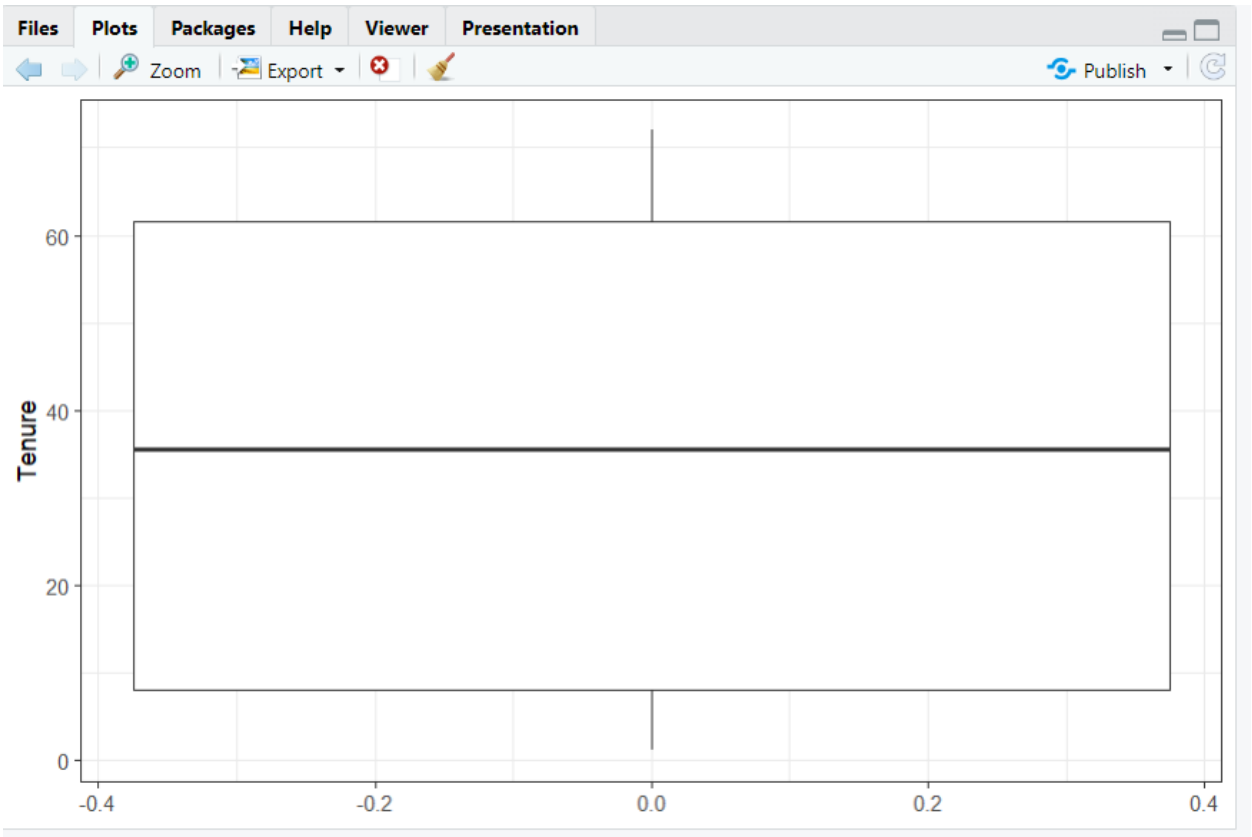

Histogram of My_data$Timely_Response
Frequency
0
500
1500
2500
3500
1
2
3
4
5
6
7
My_data$Timely_Response

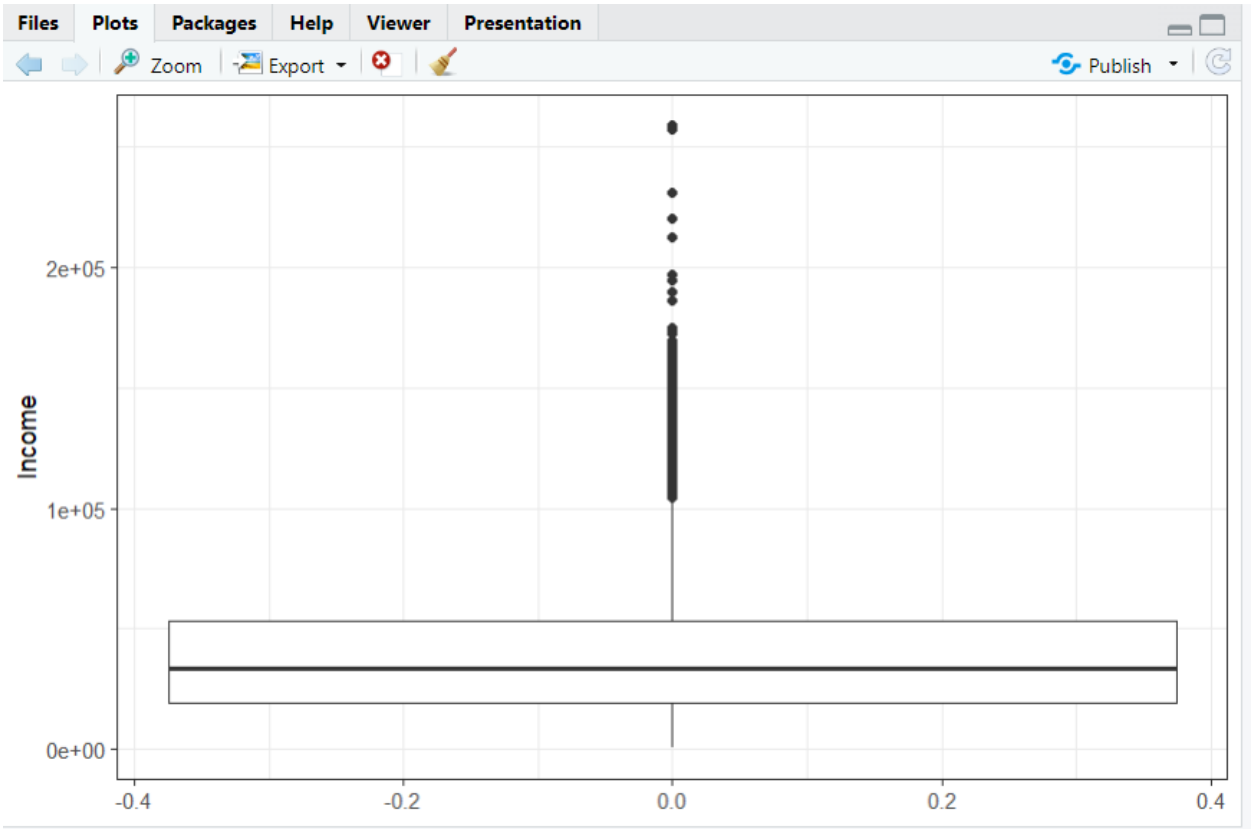

Histogram of My_data$Timely_Fixes
Frequency
0
500
1500
2500
3500
1
2
3
4
5
6
7
My_data$Timely_Fixes


Histogram of churn_clean$Income
Frequency
0
500
1500
2500
0
50000
100000
150000
200000
250000
churn_clean$Income


Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
Timely_Response
6
4
2
-0.4
-0.2
0.0
0.2
0.4

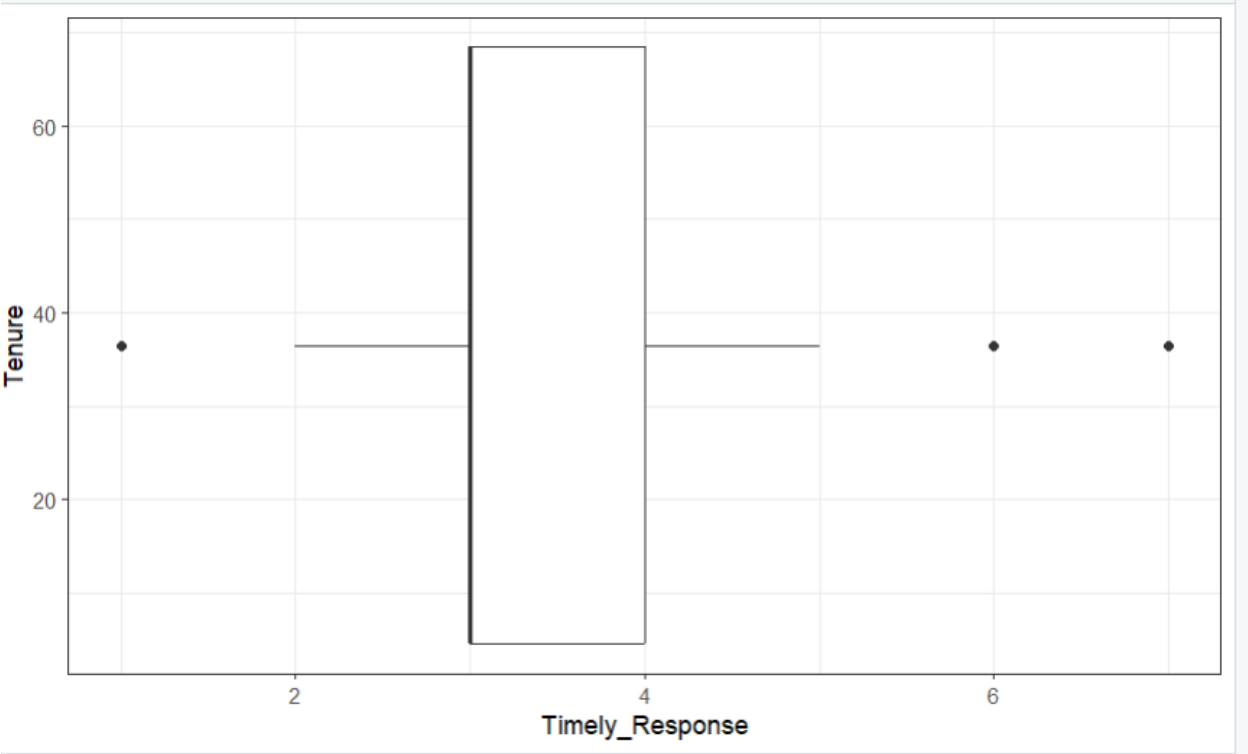

Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
Timely_Fixes
6
4
2
-0.4
-0.2
0.0
0.2
0.4

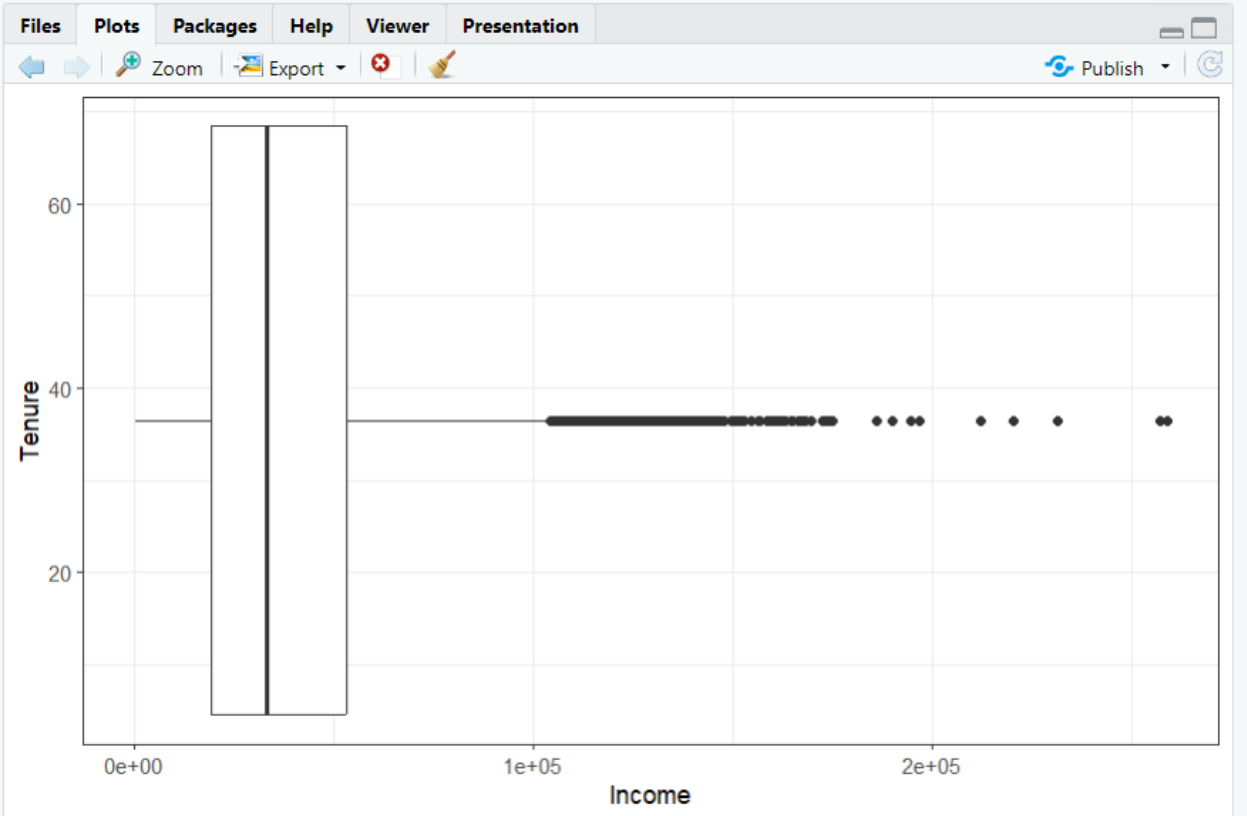

Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
Tenure
60
40
20
0
-0.4
-0.2
0.0
0.2
0.4

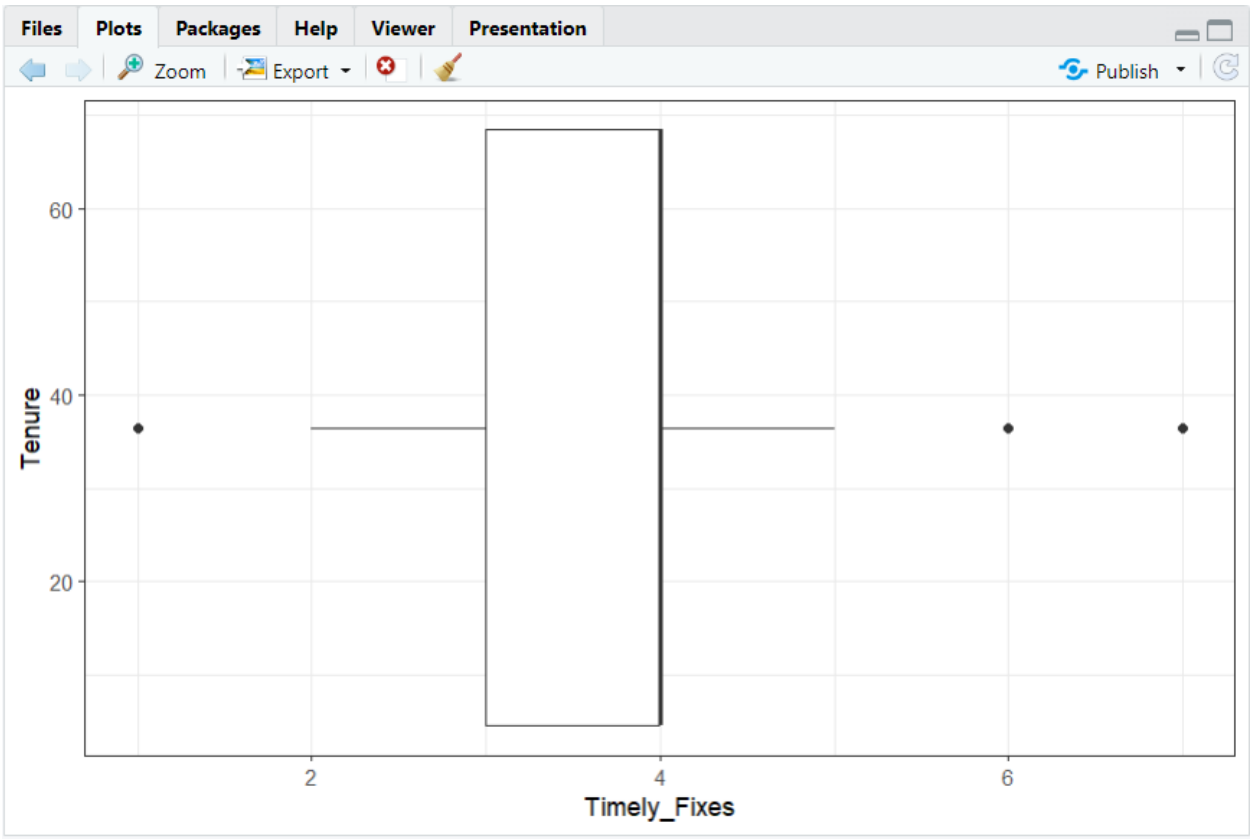

Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
Income
2e+05
1e+05
0e+00
-0.4
-0.2
0.0
0.2
0.4

```
  theme_bw()

My_data %>%
  drop_na(Tenure) %>%
  ggplot(aes(Tenure)) +
  geom_boxplot() +
  coord_flip() +
  theme_bw()

churn_clean %>%
  drop_na(Income) %>%
  ggplot(aes(Income)) +
  geom_boxplot() +
  coord_flip() +
  theme_bw()
```

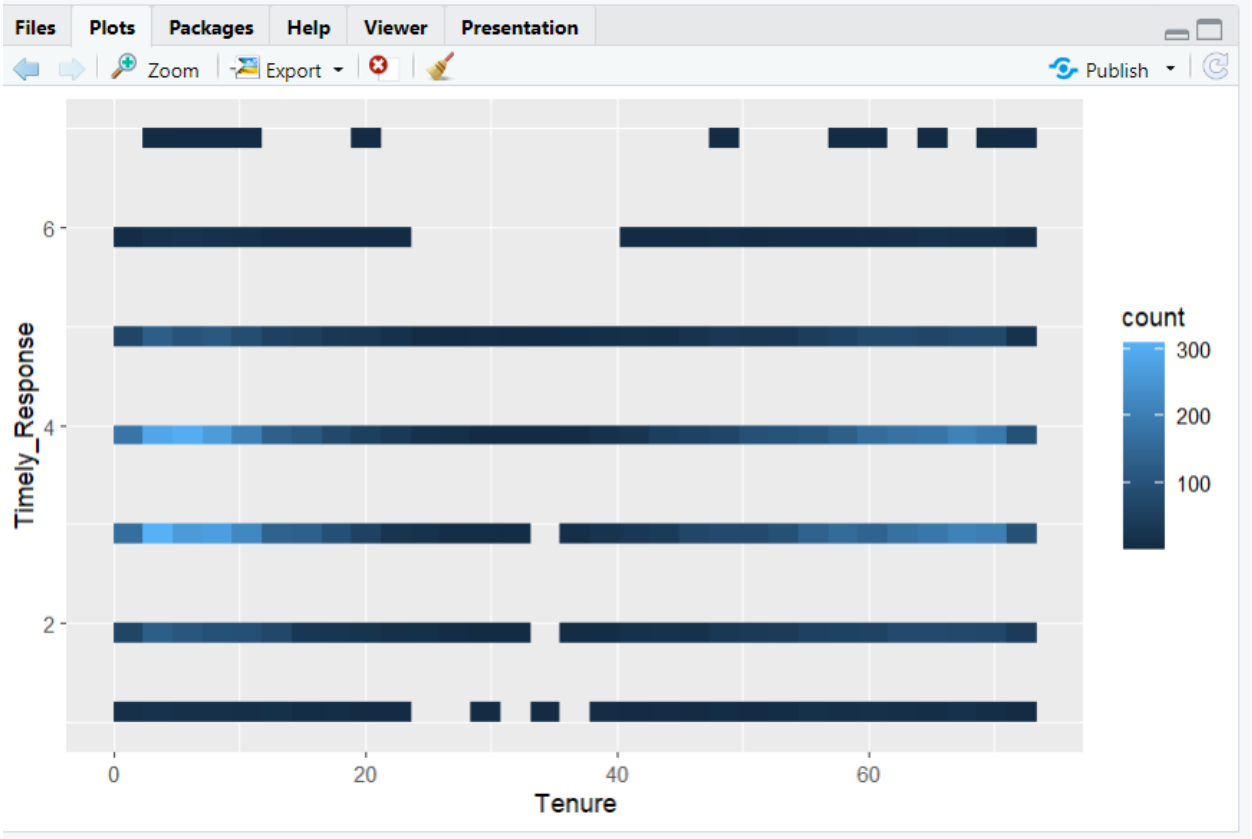
```
#Bivariate graphs

My_data %>%
  drop_na(Timely_Response) %>%
  ggplot(aes(Tenure, Timely_Response)) +
  geom_boxplot() +
  coord_flip() +
  theme_bw()

My_data %>%
  drop_na(Timely_Fixes) %>%
  ggplot(aes(Tenure, Timely_Fixes)) +
  geom_boxplot() +
  coord_flip() +
  theme_bw()

churn_clean %>%
  drop_na(Income) %>%
  ggplot(aes(Tenure, Income)) +
  geom_boxplot() +
```

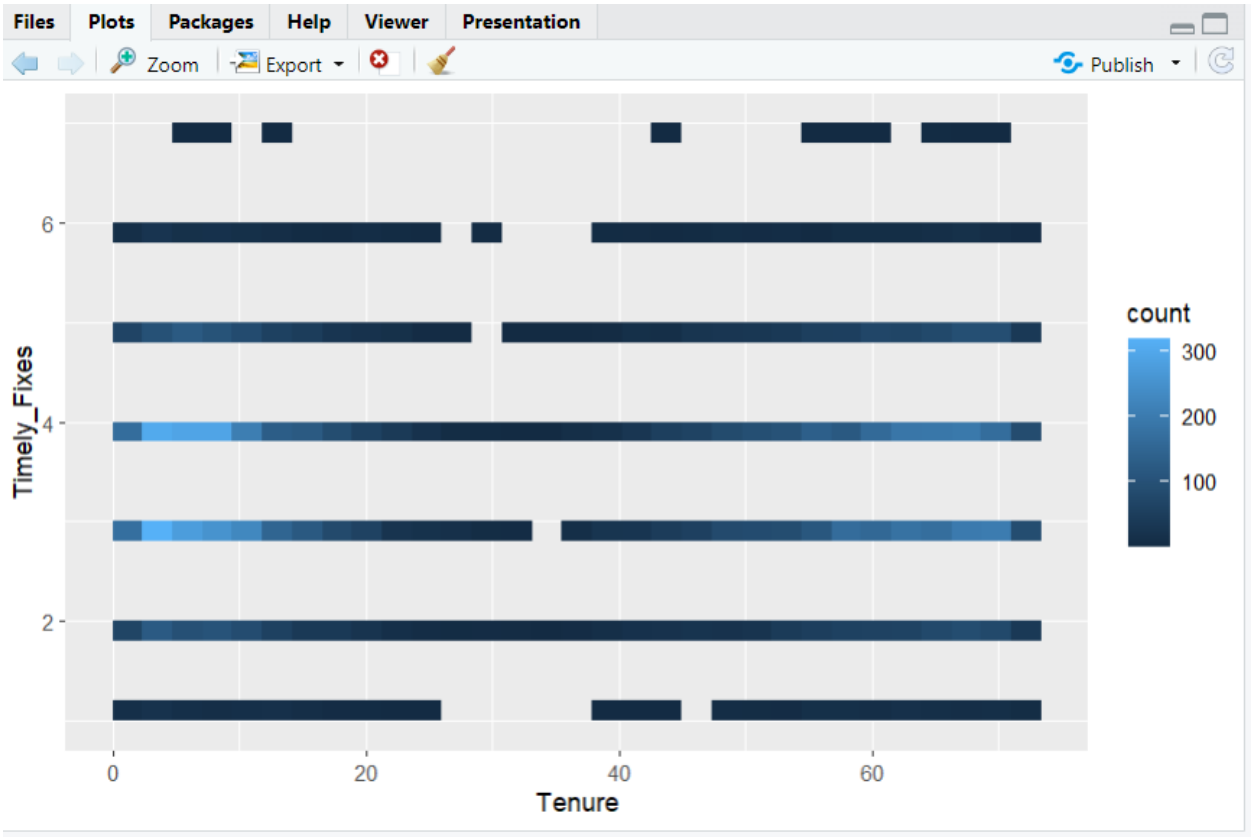

60
40
20
Tenure
2
4
6
Timely_Response


Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
60
40
20
Tenure
0e+00
1e+05
2e+05
Income




```
141
142 churn_clean %>%
143   drop_na(Income) %>%
144   ggplot(aes(Tenure, Income)) +
145   geom_boxplot() +
146   coord_flip() +
147   theme_bw()
148
149
150 #Create a heatmap
151 My_data %>%
152   drop_na(Timely_Response) %>%
153   ggplot(aes(Tenure, Timely_Response)) +
154   geom_bin2d()
155
156 My_data %>%
157   drop_na(Timely_Fixes) %>%
158   ggplot(aes(Tenure, Timely_Fixes)) +
159   geom_bin2d()
160
161
162
163
```


Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
Timely_Response
6
4
2
0
20
40
60
Tenure
count
300
200
100


Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
Timely_Fixes
6
4
2
0
20
40
60
Tenure
count
300
200
100

```
#Check working directory
	getwd()

#data profiling
	str("~/MSDA/churn_clean.csv")
	str("C:/Users/ntrei/OneDrive/Documents/MSDA/churn_clean.csv")

#dimension of churn_clean [in-text citation: (R programming 101, n.d.)]
	dim(churn_clean)
	library(tidyverse)
	glimpse(churn_clean)

#rename column names item 1- 8 [in-text citation: (Zach, 2022)]
	colnames(churn_clean)[colnames(churn_clean) == 'Item1'] <- 'Timely_Response'
	colnames(churn_clean)[colnames(churn_clean) == 'Item2'] <- 'Timely_Fixes'
	colnames(churn_clean)[colnames(churn_clean) == 'Item3'] <- 'Timely_Replacements'
	colnames(churn_clean)[colnames(churn_clean) == 'Item4'] <- 'Reliability'
	colnames(churn_clean)[colnames(churn_clean) == 'Item5'] <- 'Options'
	colnames(churn_clean)[colnames(churn_clean) == 'Item6'] <- 'Respectful_Response'
	colnames(churn_clean)[colnames(churn_clean) == 'Item7'] <- 'Courteous_Exchange'
	colnames(churn_clean)[colnames(churn_clean) == 'Item8'] <- 'Active_Listening'

#Verify columns were re-named successfully
	glimpse(churn_clean)

#explore data variables [in-text citation: (R programming 101, n.d.)]
	View(sort(table(churn_clean$Tenure)))
	View(sort(table(churn_clean$Churn)))
	View(sort(table(churn_clean$Timely_Response)))
	View(sort(table(churn_clean$Timely_Fixes)))

	barplot(sort(table(churn_clean$Churn)))
	barplot(sort(table(churn_clean$Tenure)))
	barplot(sort(table(churn_clean$Timely_Response)))
	barplot(sort(table(churn_clean$Timely_Fixes)))

	mean(churn_clean$Tenure)
	mean(churn_clean$Timely_Response)
	mean(churn_clean$Timely_Fixes)

	hist(churn_clean$Tenure)
	hist(churn_clean$Timely_Response)
	hist(churn_clean$Timely_Fixes)

	b <- boxplot(churn_clean$Tenure)
```

```
b <- boxplot(churn_clean$Timely_Response)
b <- boxplot(churn_clean$Timely_Fixes)

churn_clean %>%
 summarise(mean(Timely_Response), mean(Timely_Fixes), mean(Tenure))

churn_clean %>%
 summarise(median(Timely_Response), median(Timely_Fixes), median(Tenure))

#Create a table with data columns
My_data <- churn_clean %>%
 select(Tenure, Timely_Response, Timely_Fixes) %>%
 drop_na()

#Graph data
library(ggplot2)
 ggplot(data = My_data,
 mapping =  aes(x = Tenure, y = Timely_Response)) +
  geom_point(color = "cornflowerblue",
          alpha = .7,
          size = 2) +
  geom_smooth(method = "lm",
           se = FALSE,
           linewidth = 1.5)



#ANOVA (R programming 101, n.d.)]
#Research question: What is the effect of the survey question on the length of tenure?
#Hypothesis: Ho- there is a correlation between the mean tenure and survey question
#H1- there is not correlation between the mean tenure and the survey question
#Calculate test statistic

My_data %>%
 select(Tenure, Timely_Response, Timely_Fixes) %>%
 drop_na()

My_anova <-
 aov(My_data$Timely_Response ~ My_data$Tenure, data = My_data)

summary(my_anova)

My_anova2 <- aov(My_data$Timely_Fixes ~ My_data$Tenure, data = My_data)

summary(My_anova2)
```

See code attached.

2.

```
#ANOVA (R programming 101, n.d.)]
#Research question: What is the effect of the survey question on the length of tenure?
#Hypothesis: Ho- there is a correlation between the mean tenure and the survey question H1- there is not a correlation between the mean tenure and the survey question
#Calculate test statistic

        My_data %>%
          select(Tenure, Timely_Response, Timely_Fixes) %>%
          drop_na()

        My_anova <-
          aov(My_data$Timely_Response ~ My_data$Tenure, data = My_data)

        summary(my_anova)

        My_anova2 <- aov(My_data$Timely_Fixes ~ My_data$Tenure, data = My_data)

        summary(My_anova2)
```

```
+    aov(My_data$Timely_Response ~ My_data$Tenure, data = My_data)
> summary(my_anova)
                    Df Sum Sq Mean Sq F value Pr(>F)
churn_clean$Tenure   1      0  0.4201    0.39  0.532
Residuals         9998  10769  1.0771
```

```
> My_anova2 <- aov(My_data$Timely_Fixes ~ My_data$Tenure, data = My_data)
> summary(My_anova2)
                Df Sum Sq Mean Sq F value Pr(>F)
My_data$Tenure   1      0  0.1011   0.094  0.759
Residuals     9998  10704  1.0706
```

3.

I chose to use ANOVA as an analysis technique to answer the research question, “What effect does Timely_Response have on Tenure,” because ANOVA is a comparison test in which the effect a categorical variable has on the mean of some other characteristic. ANOVA was the test of choice because we are testing a categorical variable and a numerical variable against one another. There are three assumptions that must be made to use ANOVA. These assumptions are

that they are from normally distributed populations, equal variances, and observations are independent of one another (Zach, 2019). Timely_Response and Tenure are independent variables. Timely_Response as seen in the histogram above has a normal distribution and Tenure has a normal bimodal distribution as seen in the histogram above. The categorical ordinal variable is Timely_Response. The quantitative continuous variable is Tenure.

C.

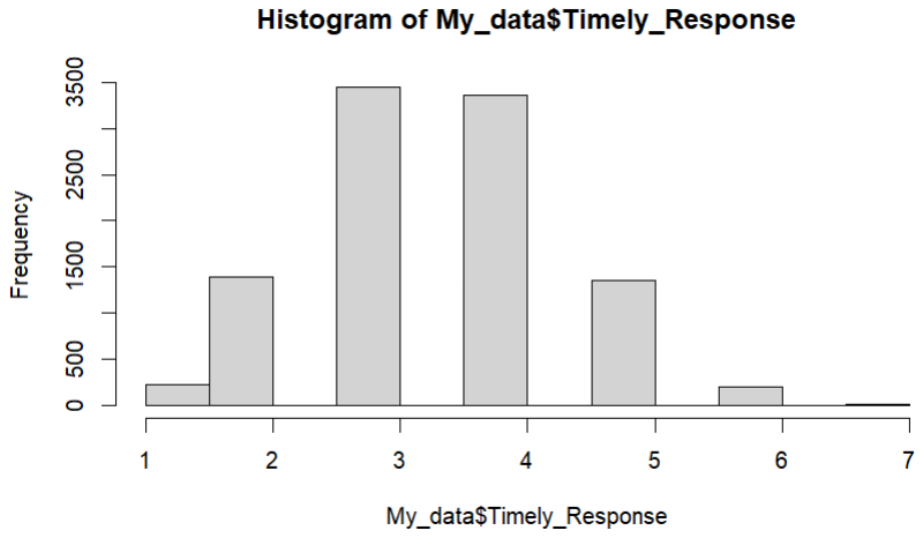
Two continuous variables are Tenure and Income. Two categorical variables are Timely_Response and Timely_Fixes.

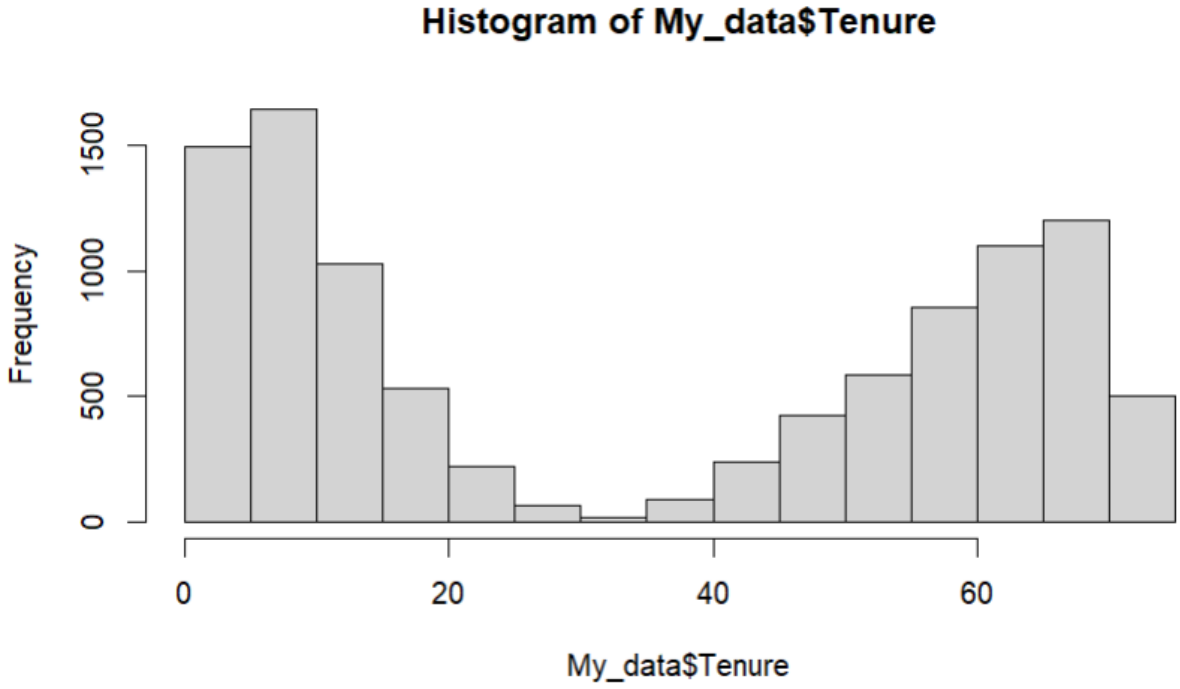
1.

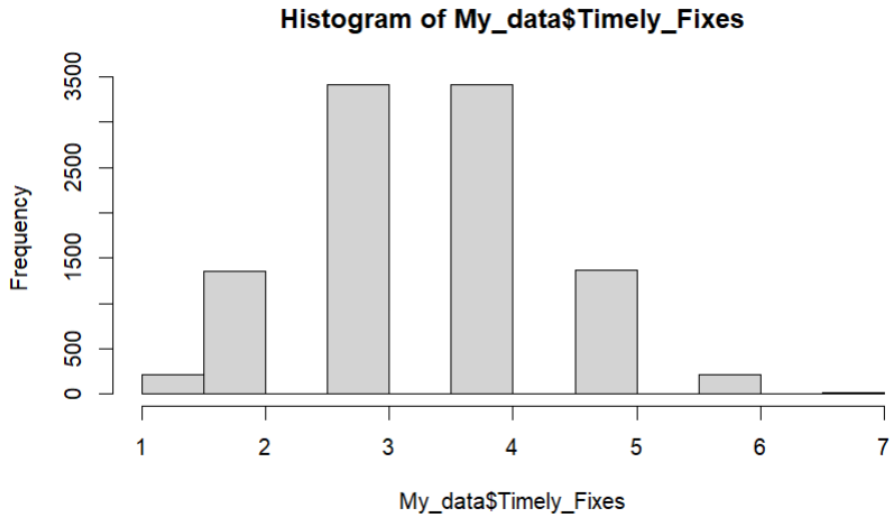
```
+    aov(My_data$Timely_Response ~ My_data$Tenure, data = My_data)
> summary(my_anova)
                     Df Sum Sq Mean Sq F value Pr(>F)
churn_clean$Tenure    1      0  0.4201    0.39  0.532
Residuals          9998  10769  1.0771
```

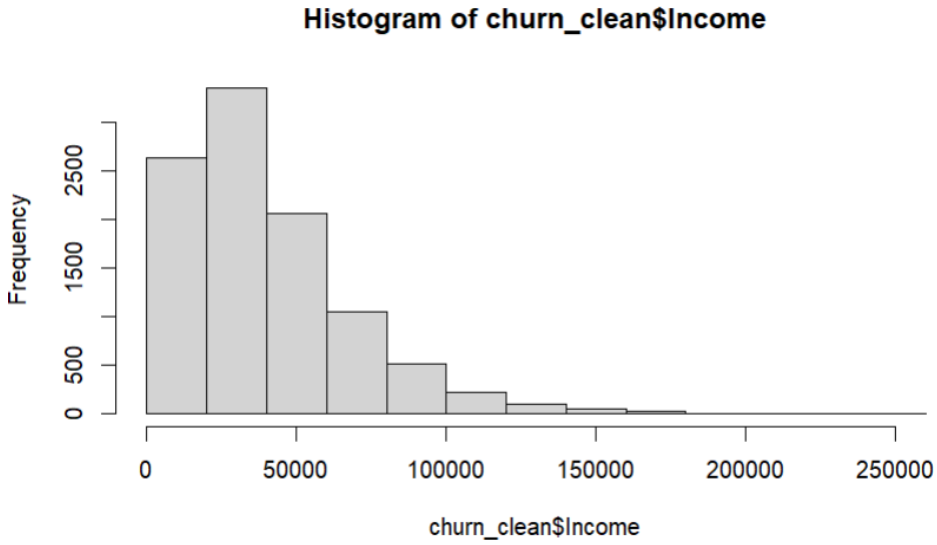
```
> My_anova2 <- aov(My_data$Timely_Fixes ~ My_data$Tenure, data = My_data)
> summary(My_anova2)
                 Df Sum Sq Mean Sq F value Pr(>F)
My_data$Tenure    1      0  0.1011   0.094  0.759
Residuals      9998  10704  1.0706
```

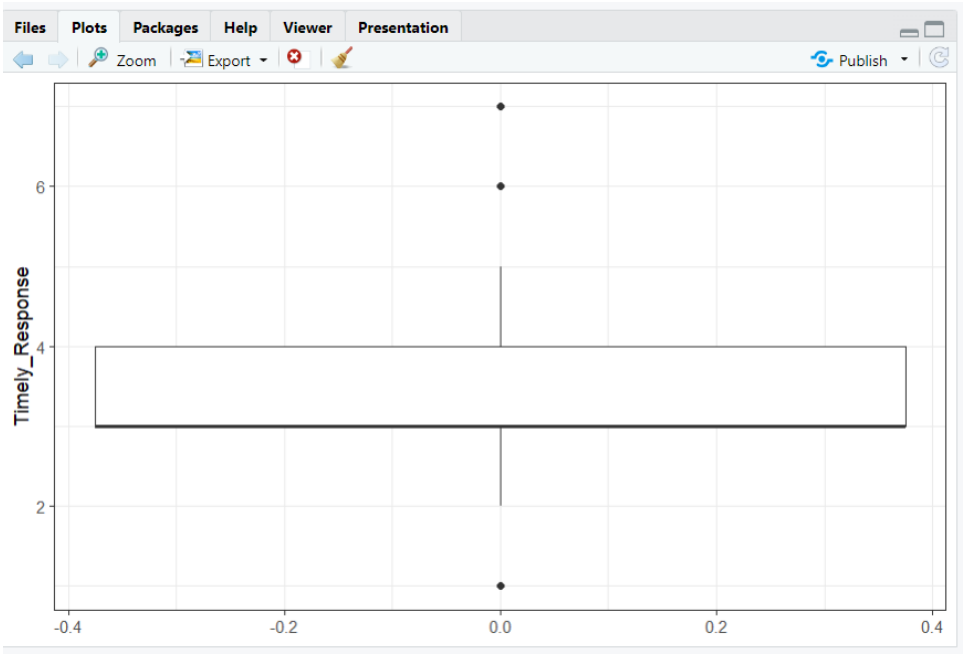
```
#univariate graphs
        hist(My_data$Tenure)
        hist(My_data$Timely_Response)
        hist(My_data$Timely_Fixes)
        hist(churn_clean$Income)

        b <- boxplot(My_data$Tenure)
        b <- boxplot(churn_clean$Timely_Response)
        b <- boxplot(churn_clean$Timely_Fixes)
```

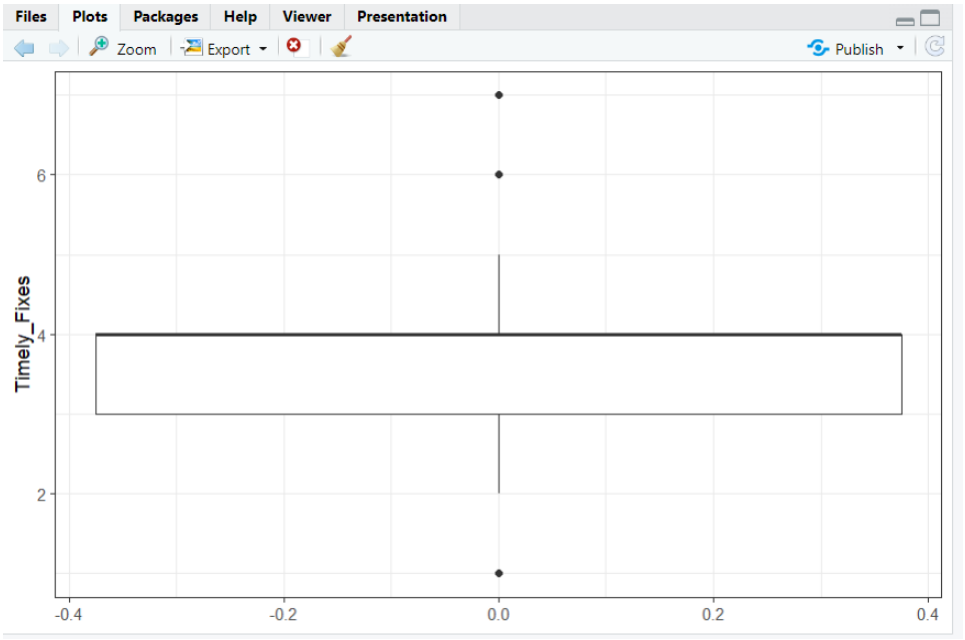

Histogram of My_data$Timely_Response
Frequency
0
500
1500
2500
3500
1
2
3
4
5
6
7
My_data$Timely_Response

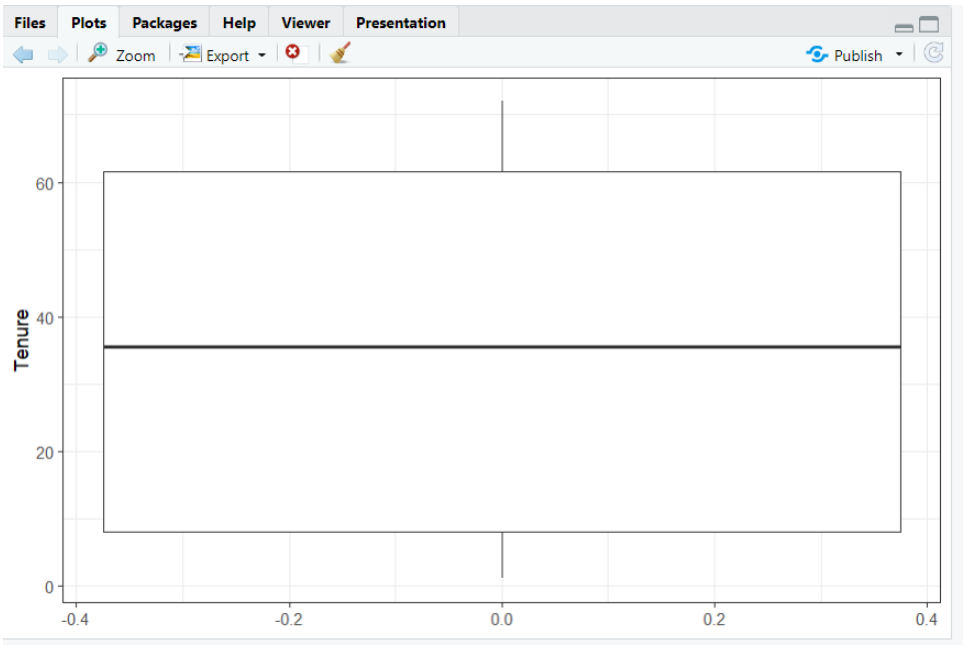

Histogram of My_data$Tenure
Frequency
0
500
1000
1500
0
20
40
60
My_data$Tenure

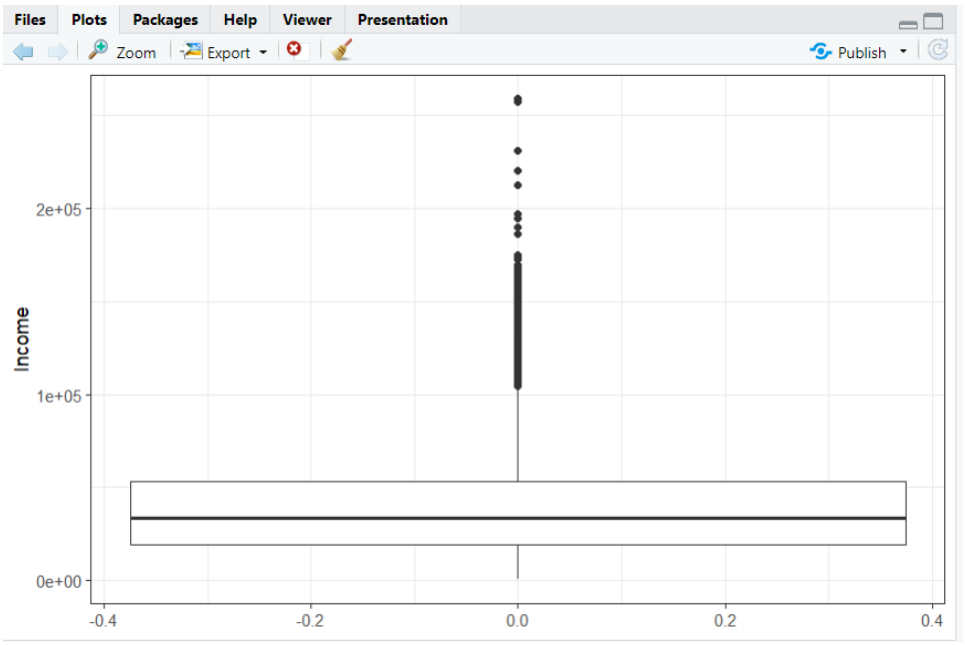

Histogram of My_data$Timely_Fixes
Frequency
0
500
1500
2500
3500
1
2
3
4
5
6
7
My_data$Timely_Fixes

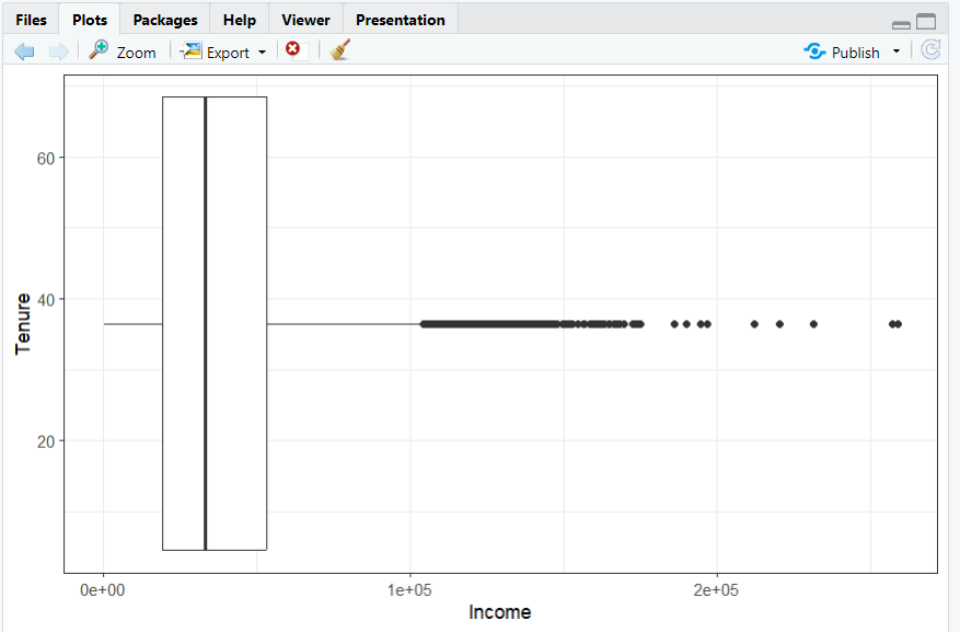

Histogram of churn_clean$Income
Frequency
0
500
1500
2500
0
50000
100000
150000
200000
250000
churn_clean$Income


Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
Timely_Response
6
4
2
-0.4
-0.2
0.0
0.2
0.4


Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
Timely_Fixes
6
4
2
-0.4
-0.2
0.0
0.2
0.4


Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
Tenure
60
40
20
0
-0.4
-0.2
0.0
0.2
0.4


Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
Income
2e+05
1e+05
0e+00
-0.4
-0.2
0.0
0.2
0.4



D.

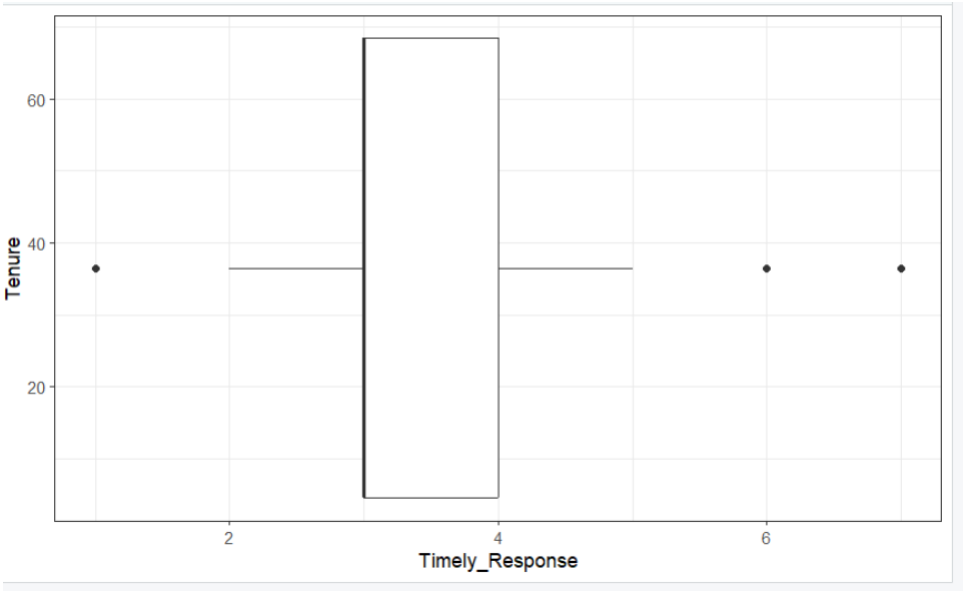
1. Two continuous variables are Tenure and Income. Two categorical variables are Timely_Response and Timely_Fixes.

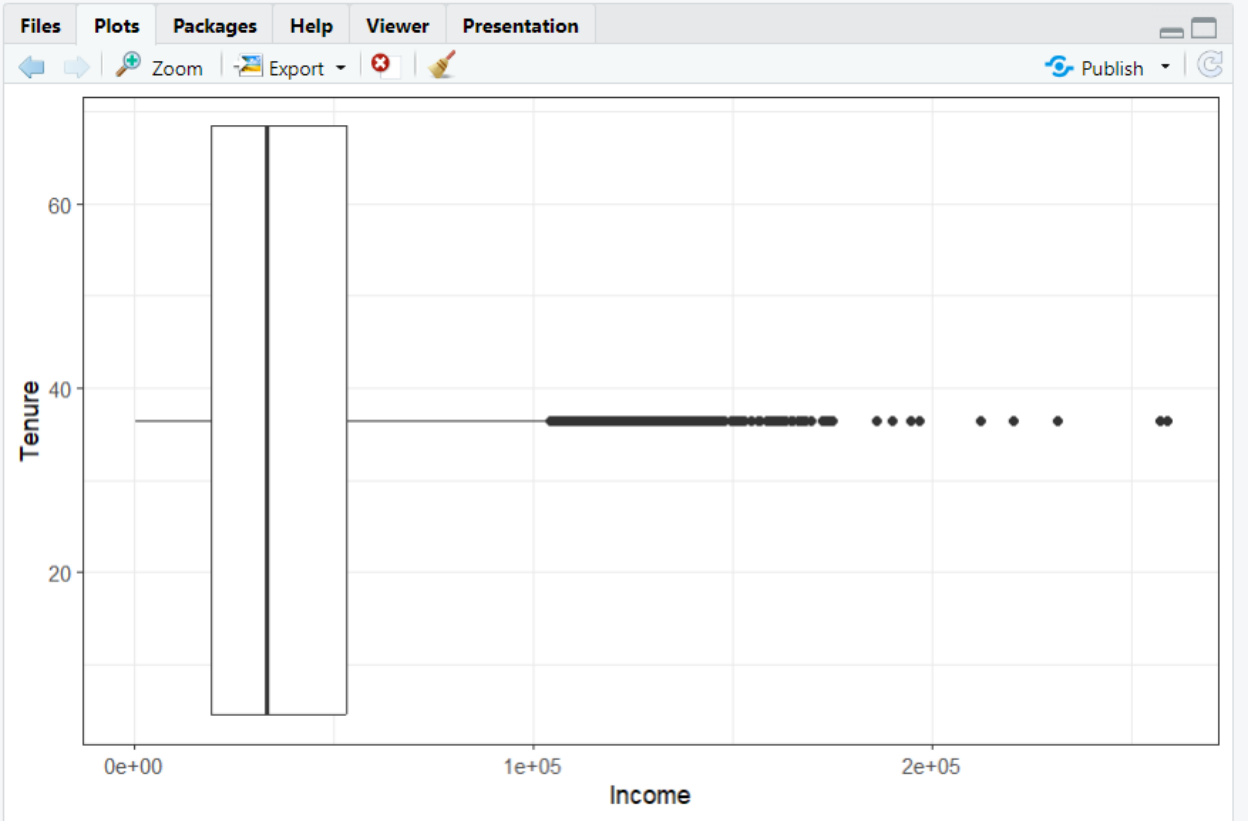
```
#Bivariate graphs
My_data %>%
  drop_na(Timely_Response) %>%
  ggplot(aes(Tenure, Timely_Response)) +
  geom_boxplot() +
  coord_flip() +
  theme_bw()


My_data %>%
  drop_na(Timely_Fixes) %>%
  ggplot(aes(Tenure, Timely_Fixes)) +
```

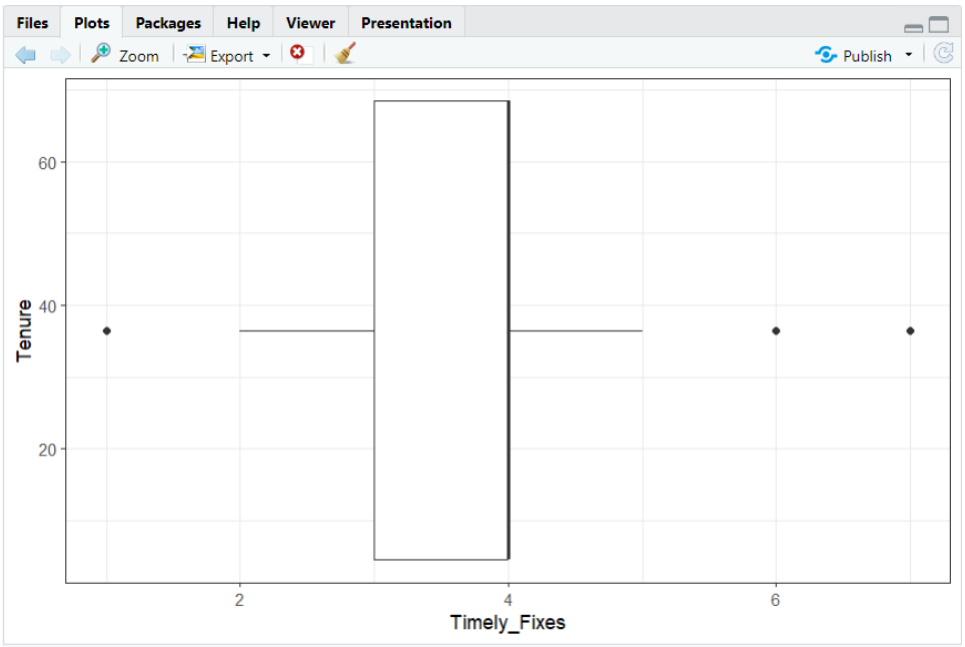
```
  geom_boxplot() +
  coord_flip() +
  theme_bw()

churn_clean %>%
  drop_na(Income) %>%
  ggplot(aes(Tenure, Income)) +
  geom_boxplot() +
  coord_flip() +
  theme_bw()



#Create a heatmap
My_data %>%
  drop_na(Timely_Response) %>%
  ggplot(aes(Tenure, Timely_Response)) +
  geom_bin2d()

My_data %>%
  drop_na(Timely_Fixes) %>%
  ggplot(aes(Tenure, Timely_Fixes)) +
  geom_bin2d()
```

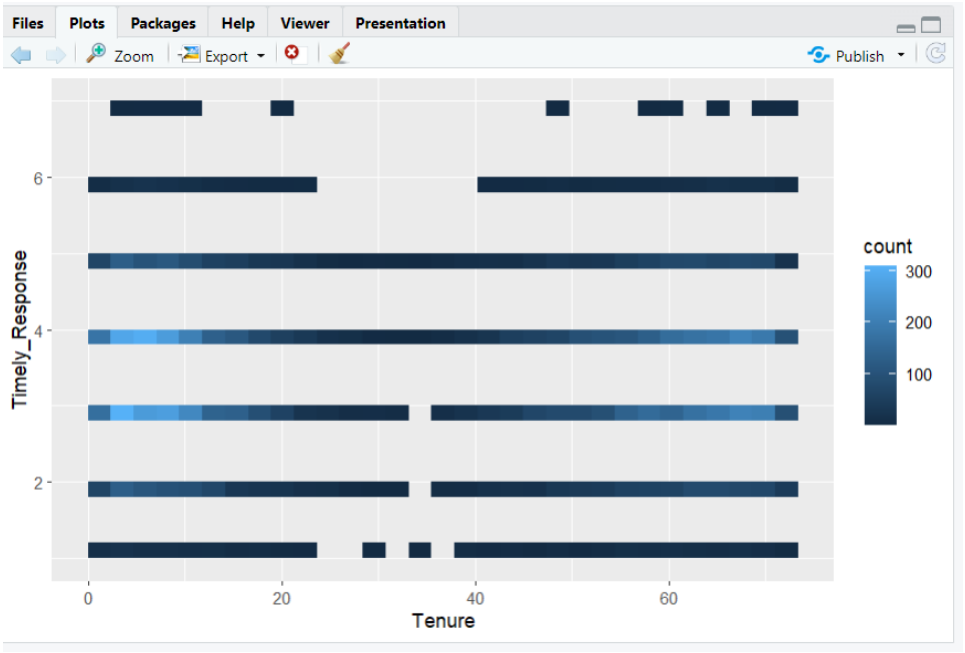
```
126 #Bivariate graphs
127
128 My_data %>%
129   drop_na(Timely_Response) %>%
130   ggplot(aes(Tenure, Timely_Response)) +
131   geom_boxplot() +
132   coord_flip() +
133   theme_bw()
134
135 My_data %>%
136   drop_na(Timely_Fixes) %>%
137   ggplot(aes(Tenure, Timely_Fixes)) +
138   geom_boxplot() +
139   coord_flip() +
140   theme_bw()
141
142 churn_clean %>%
143   drop_na(Income) %>%
144   ggplot(aes(Tenure, Income)) +
145   geom_boxplot() +
```

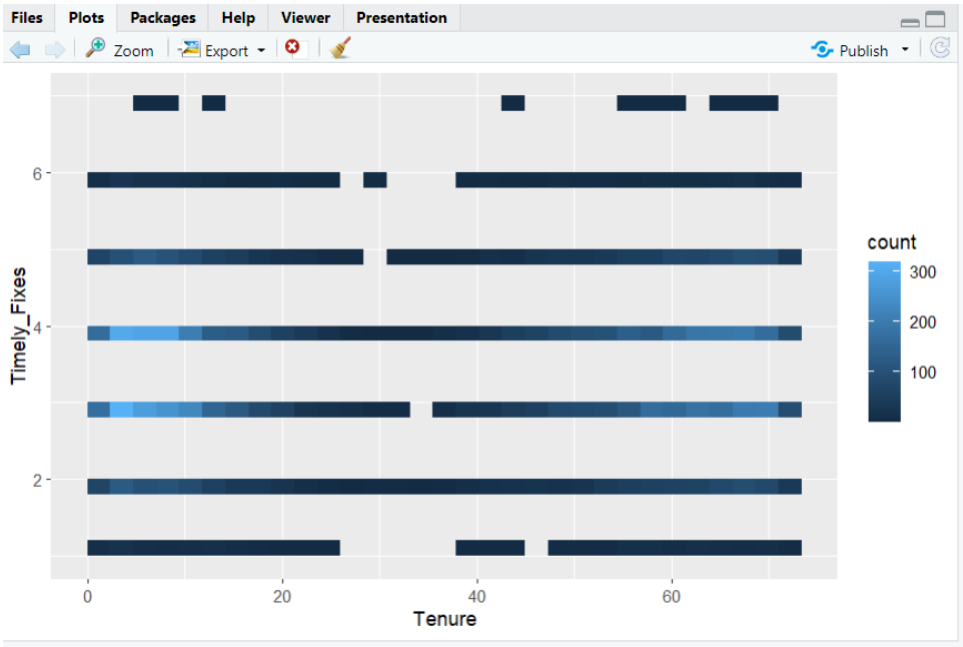
```
141
142 churn_clean %>%
143   drop_na(Income) %>%
144   ggplot(aes(Tenure, Income)) +
145   geom_boxplot() +
146   coord_flip() +
147   theme_bw()
148
149
150 #Create a heatmap
151 My_data %>%
152   drop_na(Timely_Response) %>%
153   ggplot(aes(Tenure, Timely_Response)) +
154   geom_bin2d()
155
156 My_data %>%
157   drop_na(Timely_Fixes) %>%
158   ggplot(aes(Tenure, Timely_Fixes)) +
159   geom_bin2d()
160
161
162
163
```


Tenure
60
40
20
2
4
6
Timely_Response


Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
Tenure
60
40
20
0e+00
1e+05
2e+05
Income


Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
60
40
20
Tenure
2
4
6
Timely_Fixes





E. Summarize the implications of your data analysis by doing the following:

1. The results of the hypothesis testing and analysis produce a high p-value of 0.532. The p-value tells us whether there is a difference between the variables that is statistically significant or if it is due to chance. A high p-value such as this means it is not statistically significant. There is

strong evidence of the null hypothesis since it is more than .05. Therefore, we must accept the null hypothesis and reject the alternative hypothesis.

#Research question: What is the effect of Timely_Response on the length of tenure?

#Hypothesis: Ho- there is a correlation between the mean tenure and the survey question H1- there is not a correlation between the mean tenure and the survey question

2. This analysis utilized minimal variables which is a limitation of the analysis. Having a high p-value indicates we should analyze additional survey questions and compare them to one another to determine which survey question has the highest impact on customer tenure.

3. A recommended course of action is to analyze all survey response variables and compare which ones have the most significance on length of tenure.

F. See attached Panopto recording.

G. References:

Bevans, R. (2020). *Choosing the right statistical test.*

Choosing the Right Statistical Test | Types & Examples (scribbr.com)

Donovan, K. (2019). *Data analysis and processing with R based on IBIS data.*

6 Working with Tables in R | Data Analysis and Processing with R based on IBIS data (bookdown.org)

R Programming 101. (n.d.*). T-test and interpretting p values using R programming.*

t-test and interpreting p values using R Programming - YouTube

R Programming 101. (n.d.). *Anova using R programming.*

ANOVA using R programming - YouTube

R Programming 101. (n.d.). *Statistics made easy.*

Statistics made easy ! ! ! Learn about the t-test, the chi square test, the p value and more - YouTube

R Programming 101. (n.d.). *Ggplot for graphs.*

ggplot for plots and graphs. An introduction to data visualization using R programming - YouTube

Zach. (2022). *How to rename a single column in R.*

How to Rename a Single Column in R (With Examples) - Statology

H. Sources:

Zach, (2019). *How to check Anova assumptions.*

How to Check ANOVA Assumptions - Statology